National College of Ireland

Bi-Partisan And Broader Political Spectrum Bias Detection using Transformer Based LLM Models

MSc Research Project
Master of Science Artificial Intelligence

Abin Abraham
Student ID: X23351799

School of Computing
National College of Ireland

Supervisor: Abdul Razzaq

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Abin Binoy Abraham		
Student ID:	X23351799		
Programme:	MSc. Artificial Intelligence	**Year:**	2024-2025
Module:	Practicum 2		
Supervisor:	Abdul Razzaq		
Submission Due Date:	15th September 2025		
Project Title:	Bi-Partisan And Broader Political Spectrum Bias Detection using Transformer Based LLM Models		
Word Count:	7500	**Page Count:**	24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Abin Binoy Abraham (x23351799)

Date: 15th September 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Bi-Partisan And Broader Political Spectrum Bias Detection using Transformer Based LLM Models

Abin Binoy Abraham
X23351799

Abstract

Political biasness is a very common but a concerning issue that affects the overall neutrality and unbiased reporting of facts by different media houses. The popularity of social media platforms has allowed individual users and politicians to post information that can be fake as well highly biased towards a particular political ideology. This entails a broader political polarization of common masses that can result in undermining the overall democratic discourse. To address this, the thesis focuses on implementing AI techniques to measure bi-partisan political biasness in politically charged statements as well as to measure the political biasness in a much larger spectrum by headlines of news articles published by various media houses. To implement a solution to this Transformer based LLM models like BERT, RoBERTa and DeBERTa is used for fine tuning on three datasets; LIAR, UK News and QBias datasets to generate sentence embeddings for downstream classification. This model is compared with the baseline sentence embedding models like BG-BASE-EN to evaluate the performance. The proposed methodology outperformed all the baseline models in bi-partisan political bias classification as well as classification on a broader political spectrum. The BERT fine-tuned model outperformed the baseline models for bi-partisan classification by achieving a macro F1-score of 94% which is 20% better than the baseline model.

Keywords- Political Bias Detection, LLMs, BERT, RoBERTa, DeBERTa, Sentence Embeddings, MLP

1 Introduction

1.1 Prevelant of Political Bias in Digital and mass media

Bias is an innate human trait. Unconscious bias leads to creating stereotypes that can lead to oppression and discrimination. Bias can be of different types, it could be racial bias, gender bias, affinity bias, conformity bias. But in the long term, all kinds of biasness have ill effects and can affect the harmony of a peaceful society. On the topic of bias, partisan bias is a very common type of bias, which has been a topic of concern during various democratic elections. In a general term, political bias is the inclination of media outlets, journalists, individuals, public social media platforms to favour one political party's ideology over the others. Political bias in media is a very complex issue, with a survey showing that around 80% of Americans find that there is a great deal of biasness in the media they consumed[1]. The media houses can distort or intentionally omit some facts in their reporting in order for the statement, headline or

[1] Uwgb.edu. (2019). *UW-Green Bay Library: Identifying Bias: Politics and the Media*. [online] Available at: https://library.uwgb.edu/c.php?g=600365&p=4157309.

the entire news remain favourable to their political leaning. Biased news has a tendency to manipulate as well as further polarize the individuals and the society to a political ideology. This can result in the public to be more accepting of much extreme views and opinions which was normally considered to be discriminatory and oppressive. The other way political biasness is circulated is through social media. Social media allows people from all over the world and of all classes to post any content they want. These platforms are especially used my party officials, senate or house representatives, and other organizations to interact with their voters and supporters. Many of the posts created and shared by these political entities are not curated and can also be highly biased and sometimes outright fake. This results in spreading of misinformation with a polarizing effect[2]. The other interesting part of political bias is that political parties tend to have ideology which are generally racist, misogynistic and to say the least discriminatory within its own context. For example, National Party which is a far-right political party in Ireland has many racist and discriminatory views as their core principles, therefore content that polarizes the people to sympathize towards National Party will lead them to hold discriminatory which they originally wouldn't[3].

1.2 Contribution of AI in detecting political bias

The developments in Artificial Intelligence have allowed to design and develop systems that can classify and predict the sentiments of human emotions and behaviour through texts. This approach can be extended to develop models that can analyze the statements, headlines, and texts to understand on what side of the political spectrum the mentioned content lies. Having such systems that can classify the news articles and social media content on the basis of political spectrum can help the readers and the platform users to be aware of the political biasness in the content they consume and be much for informed readers. For example; (Menzner and Leidner, 2024) presented an approach for media bias detection that moves beyond binary classification to bias subcategorization and strength estimation by presenting a fine-grained taxonomy of 27 bias types by fine tuning on GPT-3.5 and GPT-4. Additionally, (Rönnback, Emmery and Brighton, 2025) analyzed multiple form of biasness and also assess the political leanings of multiple websites using machine learning techniques. In their paper (Zhu, Cukier and Cruickshank, 2025) they introduced DocNet which is a lightweight inductive framework for political bias detection which doesn't rely on conventional pre-trained embeddings, the mode encoded the semantic structure of the documents by converting news articles to word co-occurrence graphs.

1.3 Research Questions

The thesis' is addressing two different research questions but within a similar context

[2] The Observatory on Social Media at IU Bloomington (2024). *Bias on Twitter Comes from users, Not platform, Says IU Study*. [online] Research Impact. Available at: https://research.impact.iu.edu/key-areas/social-sciences/stories/social-media-platform-bias.html.

[3] The National Party. (2018). *National Party Principles - The National Party*. [online] Available at: https://nationalparty.ie/principles/.

RESEARCH QUESTION 1: How effectively can transformer based LLM models like BERT, RoBERTa and DeBERTa can predict and classify bi-partisan political biasness in politically charged claims?

RESEARCH QUESTION 2: How effectively can transformer based LLM models like BERT, RoBERTa and DeBERTa can predict and classify political biasness in news headlines on a much broader political spectrum?

The first research question which is a binary classification problem is aimed to assess how effectively can LLM models be used to predict whether a particular statement is left or right. For example, Donald Trump a member of Republican party, which lies on the right side of the political spectrum has made a statement "I will build a great, great wall on our southern border.". How effectively can the proposed model be able to predict this statement to be leaning politically right. The second research question extends to encompass a much larger political spectrum, which makes it a multi-class classification problem. Since media outlets are known to have some kind of biasness, this will try to identify towards what spectrum the headlines try to lean on. This can include centre, right-centre, left-centre, left and right.

To assess this, the thesis is aiming to build three models which are fine-tuned using transformer based LLM models; BERT, RoBERTa and DeBERTa on three different datasets to generate embeddings and then using the generated embeddings as an input for a downstream classifier; which in this case is a 6-layer multilayer perceptron MLP. Tokenization is performed on the text columns of the datasets before using it for fine-tuning. Mean pooling method is applied to generate final embeddings from each of these models. This proposed model is then compared with three other baseline models. The baseline models are created using state-of-the-art lightweight sentence embedding language models such as BGE-BASE-EN, E5-BASE-EN and MPNET-BASE-V2. The model's performance is then evaluated using accuracy, macro F1-score along with weighted and micro F1-score and AUC-ROC score for the first research question, since it is a binary classification problem.

The results indicated that the fine-tuned LLM models each outperformed the baseline models in terms of performance for the first research question. BERT, RoBERTa and DeBERTa models each achieved a macro F1-score of 94%, 92% and 93% respectively indicating that BERT is the best performing model for this task. While the proposed models did outperform the baseline models for multi-class classification task, it did struggled a lot to achieve a decent overall performance, this was mainly due to high class imbalance. The models achieved a macro-F1 score of 54%(BERT), 63%(RoBERTa), 55%(DeBERTa) for UK News dataset, even though all these models maintained an accuracy of 70% or above on the same dataset. While for QBias dataset, all the models performed poorly in terms of both F1-score as well as accuracy.

The rest of the report contains, a review of related work, a detailed description of the methodology, followed by evaluation, discussion and evaluation.

2 Related Work

Transformer models like BERT are extensively used because of their ability to understand and process word sequences by focusing on the important parts of the input. Semantic models like BERT places similar meaning words close together in a conceptual space; this allows in handling of long sentences and complex language structures in an efficient manner. Due to this; transformers are heavily used in the field where semantic understanding is important, this includes in designing processes designed for detecting fake news.

2.1 Transformer models for Fake News Detection

To showcase how transformer models perform better than baseline Hybrid CNN models in the detecting fake news, (Qazi, Khan and Ali, 2020) proposed a Transformer architecture to improve the performance of fake news classification using textual claims from LIAR dataset. In their work, the positional encoding enabled the model to learn word order which is essential to determine the tone or intent of a claim. Their experiment outperformed the baseline CNN model with a validation accuracy of 39%. Their work has not been explored transformer based large language models like BERT. The capability of BERT to detect fake news was explored by (Pavlov and Mirceva, 2022) in their paper, implemented and fine-tuned two transformer models, BERT and RoBERTa to detect COVID-19 related fake news using ConstraintAI's 2021 COVID-19 fake news dataset. After individually fine tuning the model on the dataset for prediction. The BERT model achieved an accuracy of 98.31% accuracy and a F1-score of 98.31% while Roberta model achieved a F1-score of 97.52%. This showcases that Transformer based LLM models can be very effective when dealing with semantic context.

2.2 Machine Learning Approached for Political Bias Detection

The main objective of this thesis is to detect political biasness. Even though the premise is different it follows an ideal approach when detecting political biasness, since fake news and biased statements has their own way framing sentences. A recursive neural network (RNN) model to detect liberal and conservative political ideology at the sentence level was proposed by (Iyyer *et al.*, 2014). Their RNN model outperformed baselines logistic regression and word2vec-averaged classifiers. In their paper (Gangula, Duggenpudi and Mamidi, 2019) introduced Headline Attention Network to detect political bias in a custom annotated dataset. Using HAN allows for interpretable attention mapping that shows which words trigger bias. In a similar context regarding political bias detection in news articles, (Baly *et al.*, 2020) addressed the task of identifying political bias in news articles which are categorized as left, center or right. Adversarial Media Adaptation and Triplet Loss Pre-training are two debiasing strategies introduced by them. Their proposed methods achieved an accuracy of 72% and a macro-F1 of 64.29% compared to a baseline of 36.75% accuracy and 35.53% macro-F1. The performance of their model was mediocre.

2.3 Transformer based LLM models for Political Bias Detection in Media

To measure political biasness using LLM models, (V, Singh and Kumar, 2025) presented a supervised approach were where a fine-tuned Roberta model is used detecting political bias in news. To maintain the class distribution stratified sampling was implemented. Tokenization was performed using RoBERTa tokenizer and also applying truncation and attention masking. The model achieved a macro-F1 score of 94.49%. Apart from detecting political biasness in English Language (Abdelhameed, Mohamed and Medhat, 2025) proposed an ensemble model that has been trained on four datasets. FigNews, ThatiAR, SemEval-2019 Task 4 (translated), and Us Vs Them (translated). They used Arabic transformer tokenizers such AraBERT, MARBERT, and QARiB were used. Class imbalance in the dataset was handled using SMOTE, which is not a great approach since SMOTE can result in overfitting. They used a base classifier such as Random Forest and Gradient boosting which is combined with a fine-tuned transformer model. Their proposed ensemble model outperformed standalone classifiers and even prior state-of-the-art models by achieving top F1 scores across datasets: 0.8119 (SemEval), 0.5741 (ThatiAR), 0.95 (FigNews), and 0.2530 (Us Vs Them). (Krieger *et al.*, 2022) provided a strong empirical and methodological contribution to the research question on how AI models can effectively detect political bias in digital media by addressing the linguistic sensitivity of media bias and especially when it comes to word choices that results in bias. the study introduces domain-adapted transformer models like DA-RoBERTa, DA-BERT, DA-BART, and DA-T5, all pre-trained on the Wiki Neutrality Corpus (WNC) and fine-tuned on the gold-standard BABE dataset. Da-RobertA model achieved a F1-score of 0.814 which outperformed the baseline F1-score of 0.799 and thus becoming state-of-the-art.

Apart from bias detection in news headlines. the political biasness can be measured in social media posts. In their work (Gunhal *et al.*, 2022) used multiple transformer architectures for stance detection in localized policy elections. They addressed Class imbalance through back-translation data augmentation. RoBERTa-base achieved the best ternary stance detection accuracy of 84.08%. Extending the biasness on LLM models itself (Lunardi, La Barbera and Roitero, 2024) investigated the reliability of using direct questioning methods to assess the political bias in large language models and evaluated on two main consistency measures; polar consistency and paraphrastic consistency. They found that LLMs often fail to maintain consistent stances when question phrasing is altered and that it's difficult to accurately understand true political leanings of the models. (Won-Tae Joo, Jeong, and KyoJoong Oh, 2016) In their paper proposes a five-phase pipeline for detection of political bias from Korean newspapers, considering indirect expression of bias. The articles are pre-processed with KKMA for morpheme tagging, word vectors are trained with skip-gram Word2Vec, and document vectors are obtained by averaging and concatenation. (Kim *et al.*, 2023) presented a multi-stage prompt tuning framework for low-resource political perspective detectionIn stage one, domain-specific prompts are adapted via a Masked Political Phrase Prediction task on unlabeled political news to inject domain knowledge. In stage two, task-specific prompts are tuned for classification while freezing the model and domain prompts, preventing catastrophic forgetting. Tested on SemEval-2019 and AllSides, the method outperforms fine-tuning and prior prompt tuning approaches, especially in few-shot settings, with only 2.79% of parameters are trainable.

2.4 Research Gap

Most of the research is heavily focused on detecting media bias on headlines, rather than on politically charged statements. Statements are always said from a first-person view while news headlines are mostly viewed or states in a third person perspective, indicating there will be a semantic differences between news headlines and statements even if there is political bias associated with it. RQ1 is trying to answer this by performing a bi-partisan classification of politically charged statements using LIAR dataset. Other works have more focused on detecting overall biasness rather than detecting the ideological leanings of the claims or headlines. With RQ2 the thesis will try to evaluate the performance of transformer models to detect political biasness in a much larger spectrum and not focusing on just bi-partisan ideology, this larger spectrum includes left-centre, right-centre, centre.

3 Research Methodology

3.1 Dataset Description

For RQ1, the thesis will be using LIAR Dataset (Wang, 2017), which is benchmark dataset designed for fake information detection and credibility analysis of political claims. The original dataset consists of around 14,000 rows and 13 columns. In LIAR dataset, the claim column in the dataset contains the actual statement, while the label column is a rating of the truthfulness of the claim where the value ranges from TRUE to FALSE and PANTS-ON-FIRE. The dataset also contains metadata about the speaker, their political affiliation and their designated position along with an historical credibility of the speaker. The party affiliation of the dataset is mainly consists of 4 values, which are affiliated to republican, democratic party or unknown affiliations.

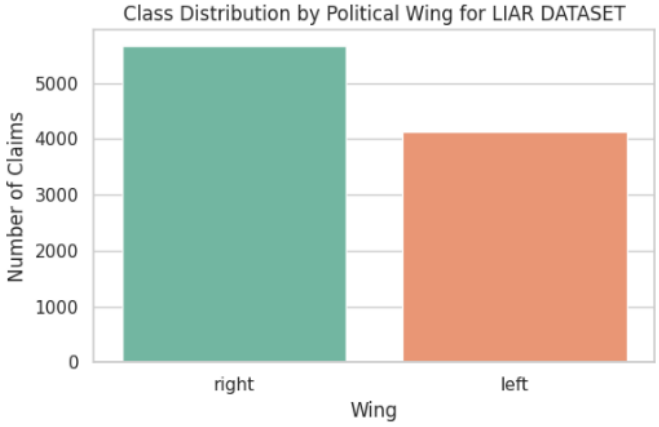


Figure 1. Class distribution in LIAR Dataset

From Figure 1. we can see there are two classes in the dataset with a slight class imbalance, making this data suitable for binary classification

The second dataset; UK News dataset is generated using web scraping method through RSS feed and is sourced from Kaggle (source:- https://www.kaggle.com/datasets/dexmasa/uk-news-headlines). The original dataset consists of around 33,000 rows and 4 columns, where website and headline are the main columns to perform the tasks. This dataset is collection of

headlines by various British based media outlets such as BBC, Daily Mail. From Figure 2. it is observed that there is a huge class imbalance in the dataset 11,276 for left-center, 7,308 entries for right, 2,815 entries for right-center and only 385 entries for center. This dataset will be used to answer the RQ2 making it a multi-class classification task to classify 4 classes; left-center, center, right-center, right.

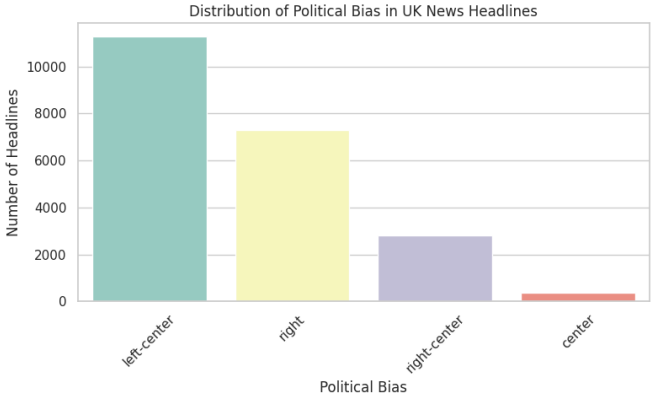


Figure 2. Class Distribution in UK News Dataset

QBias Dataset (Haak and Schaer, 2023) which is the third dataset and will be used for RQ2. This dataset is used for three class classification; left, right and centre. Originally the dataset contains 21,750 entries.

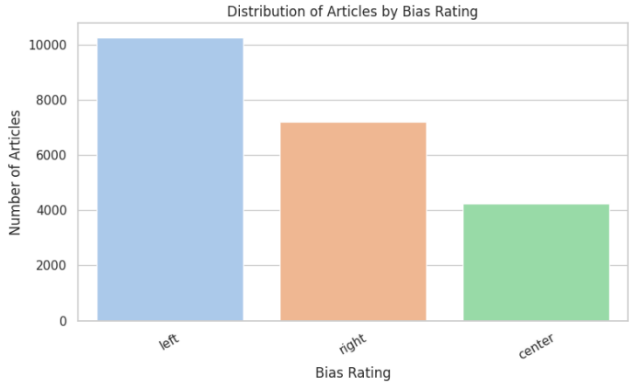


Figure 3. Class Distribution in QBias Dataset

The class distribution from Figure 3. indicates that there is a huge class imbalance where left wing is overly represented while center is in minority.

3.2 Data Preprocessing

Common preprocessing tasks will be performed on all the three datasets. This includes normalization of the text or claim related columns by converting it into lower case in case of upper-case strings. URLs in the dataset will be removed and also if there are duplicate rows with the same text content, it is removed to ensure there is no redundancy. Low quality text content, especially if they are below 10-word counts are also removed during preprocessing. Missing values will also be dropped from all the datasets.

The final Dataset records after preprocessing is given in the below table (Table 1.):

Dataset	Left	Left-Center	Center	Right-Center	Right	Total Records

LIAR(RQ1)	4,138				5,665	9,803
UK News (RQ2)		11,276	385	2,815	7,308	21,784
QBias (RQ2)	10,275	–	4,253		7,226	21,754

Table 1. Dataset Description

3.3 Baseline models using Sentence Embedding models.

Three state-of-the-art pre-trained sentence embedding language models were used; BGE-base-en (1024 dimensions) introduced in the paper by (Luo *et al.*, 2024). This embedding model is a member of BAAI (Beijing Academy of Artificial Intelligence) family and can convert any English texts to a compact vector.

In their paper, (Wang *et al.*, 2024) proposed E5-base-v2 (768 dimensions), a weekly supervised pre-training sentence embedding model.

The third baseline embedding model, MPNET-base-v2 (768 dimensions) introduced in (Song *et al.*, 2020). This embedding maps paragraphs and sentences into a dense vector space. Since this embedding model was fine-tuned on billions of datasets it is claimed to be effective in tasks which requires sentence similarity and clustering.

Above mentioned models are lightweight sentence embedding models, and they handle the tokenization of the words and final mean pooling for embedding generation within the models. Therefore, any external tokenization or mean pooling steps are not required. These models are known for capturing the semantic information to a great extent. For this thesis we will be using the pre-trained versions of these models to generate embedding and store it as .npy files and the use these embeddings as an input for downstream classifier, making it an approach similar to the work done by (Won-Tae Joo, Jeong, and KyoJoong Oh, 2016), but instead of the last stage being a regression model, it is a classification model in this thesis.

3.4 Fine-tuned Transformer models

To build fine-tune models, we will be using BERT, RoBERTa and DeBERTa. These models are used in many of the related work, for instance (V, Singh and Kumar, 2025) used RoBERTa for political bias detection.

BERT, RoBERTa, and DeBERTa are transformer models that create contextual embeddings but they vary in their architectures, training methods, and overall performance. BERT, launched in 2018, was the first to achieve bidirectional context comprehension through training on masked language modeling (MLM) and next sentence prediction (NSP). RoBERTa (2019) enhanced BERT's method by eliminating NSP, utilizing bigger datasets, and fine-tuning training parameters. DeBERTa (2020) did architectural advancements such as disentangled attention and a refined mask decoder, enhancing accuracy in representing token associations. Every model relies on its forerunner, striking a balance between efficiency, data consumption, and precision. The base versions of these models will be used in this thesis, BERT-base with around 110 million parameyers, RoBERTa-base with around 125 million parameters and DeBERTa-base with 86 million parameters

.All these models will be fine-tuned on the all the three datasets independently to answer both of the thesis' research questions. Unlike the baseline embedding approach, this model requires

tokenization of textual data as well as mean pooling to generate embeddings of 768 dimensions which will be used as input for the downstream classifier.

3.5 Downstream Classifier

The generated embeddings from both fine-tuned and baseline models will be used for downstream classification. The downstream classifier in this thesis is a 6-layer MLP. This is a deep feed-forward classifier that would initially take a fixed size embedding vector generate by the Transformer and sentence embedding models and then it maps to a logit depending on the number of classes. The input dimension for the first layer is 768 and has a hidden layer of 512. In total this MLP has 6 linear layers and ReLU activation layer is in all the layers. The dropout value after each layer is defined as 0.3. There is no softmax in this architecture since raw logits is expected by nn.CrossEntropyLoss, therefore LogSoftmax is done internally.

3.6 Political Bias Detection in Fake News/Statements

After the embeddings are generated from baseline models and fine-tuned transformer models, these embeddings for LIAR dataset, it then will be used for downstream classification using a six-layer MLP. The results of this task will be evaluated to answer the RQ1.

3.7 Political Bias Detection in News Headlines

For RQ2, the embeddings generated on UK News and QBias datasets by baseline and fine-tune models will be used for detecting political bias in the news headlines using a downstream classifier; in this case is a 6-layer MLP. This will be a multi-class classification problem predicting on what political spectrum the headline lies on.

3.8 Addressing Class Imbalance

The class imbalance in the datasets will be addressed by assigning class weights during training of the model. SMOTE and Random Oversampling approach is avoided in order to avoid overfitting.

3.9 Evaluation Metrics

Model performance was evaluated using multiple metrics. Such as Recall, Precision, Accuracy, F1-score. AUC-ROC graphs are plotted for binary classification tasks which is RQ1.
Accuracy showcases the correctly classified data to total amount of data in the dataset. While Precision displays the model's overall performance when it comes to classifying positive predictions, meanwhile is a sensitivity of the model to predict actual positive cases from overall positive predictions. F1-score is used as a key metric to detect the overall performance of the model when it comes to classification tasks. The method to calculate the following metrics is given below.

$$Accuracy = \frac{\text{True Positives} + \text{True Negatives}}{\text{True Positive} + \text{True Negative} + \text{False Positive} + \text{False Negative}}$$

$$Precision = \frac{\text{True Positives}}{\text{True Positive} + \text{False Positive}}$$

$$Recall = \frac{\text{True Positives}}{\text{True Positive} + \text{False Negative}}$$

$$F1 - score = 2 \text{ x } \frac{\text{Precision x Recall}}{\text{Precision} + \text{Recall}}$$

4 Design Specification

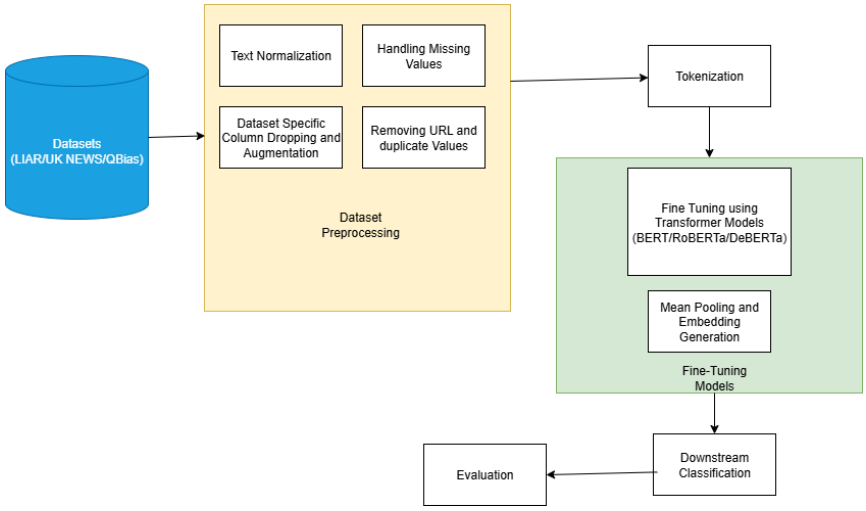


Figure 4. Design Framework

Figure 4. Shows the framework of the proposed design for the thesis.

4.1 Preprocessing

This includes Text normalization, handling missing values, removing redundant values and dataset specific augmentation like dropping columns and creating new columns using auxiliary information if required.

4.2 Tokenization

This involves tokenizing the appropriate text columns in the dataset for the transformer models to use it for fine tuning using the appropriate tokenization module.

4.3 Fine Tuning

This part deals with fine tuning the models by defining training arguments such as epochs, calculating weights and defining loss function. This also includes mean-pooling for generating final embeddings.

4.4 Downstream Classification

The generated embeddings are used as input for downstream classification using a six-layer MLP. Binary classification for LIAR dataset and multi-class classification for UK News and QBias dataset

4.5 Evaluation

The results of the classification tasks are evaluated to answer both RQ1 and RQ2.

4.6 Experiment Setup

The entire thesis practical is done on CoLab using colab's T4 GPU. Other software such as excel and google sheets are also used to explore the dataset.

5 Implementation

5.1 Dataset Specific Preprocessing Module

Figure 5. Word Cloud of LIAR Dataset

From Figure 5. word cloud image we can see that there are common terms such as "year," "state," "percent," and "people" in both the ideologies. Although left-leaning claims often emphasize topics like "republican," "job," "law," and "pay," while right-leaning claims more frequently mention "obama," "health care," "democrat," and "government."

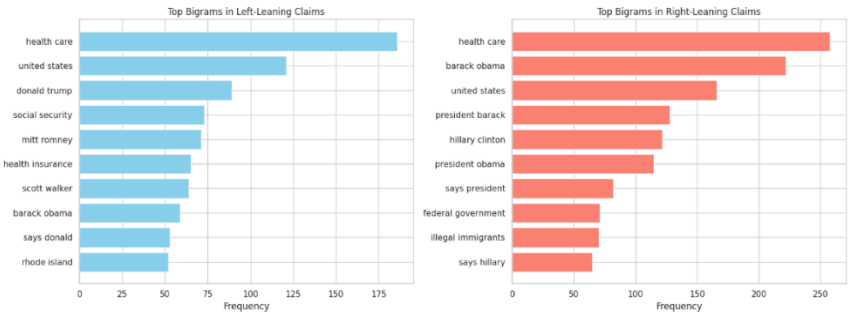


Figure 6. Bigram plot for LIAR Dataset

The bigram plots from Figure 6. reveal shared themes like "health care" and "united states," but also clear ideological differences. Left-leaning claims highlight social programs and conservative figures, while right-leaning claims focus more on liberal leaders and government critique.

The final pre-processed dataset will be used for binary classification that is to predict right or left wing claims.

For RQ2, in the UK dataset a new column called ‘political bias’ is created where the media outlets political leanings are mapped using an auxiliary data information provided in the dataset. The new column will have four unique values; right, right-center, center and left-center thus allowing to build models for multi-label classification. The preprocessing steps will be similar to that used in LIAR dataset except no prepending is performed.

Figure 7. Word Cloud for UK News Dataset

The word clouds showcase (Figure 7.) the popular terms used in headlines across the different political bias categories. Common terms such as “police”, “Nicola Bulley” appear in all classes. But there are some unique emphases for each class. Sports related terms are dominated by right-center headlines while left-center and right sources of headlines mention crime, public figures and social issues. Center aligned headlines focuses more on geopolitical topics.

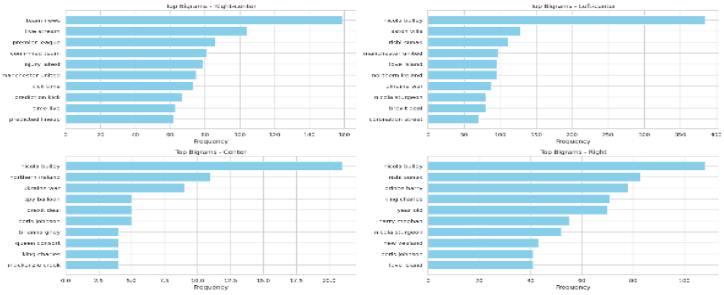


Figure 8. Bigram Plot for UK News Dataset

The bigram frequency plots as shown in Figure 8. display clear thematic differences across all the political bias categories. Right-center sources are more dominated by sports related phrases. Meanwhile left-center and right leaning headlines are more focused on public figures and events. While center-aligned outlets are focues more on political and international topics. After performing the dataset perprocessing, the entries are reduced to 21,784 rows.

Another dataset for RQ2 is QBias where the bias_rating column in QBias dataset with values left, right and center will be the label for classification.

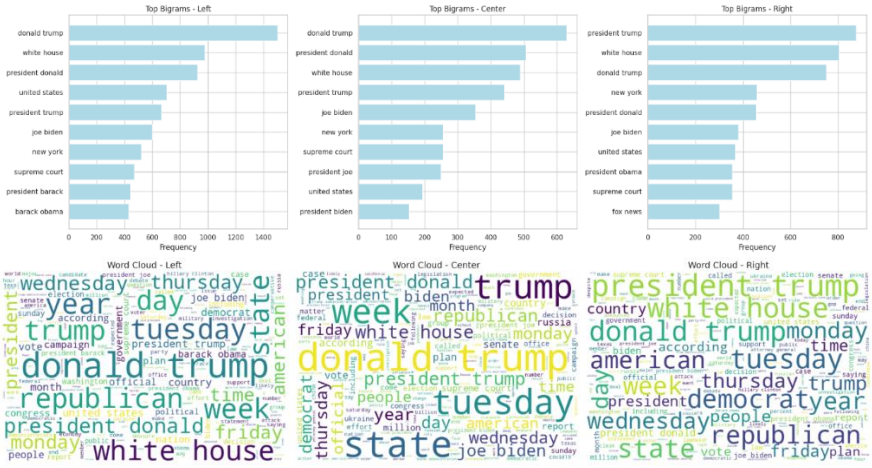


Figure 9. Word Cloud and Bigram Plot for QBias

The word cloud and bigram frequency plots as shown in Figure 9. for the QBias dataset displays subject overlaps across the political spectrum. All the bias categories are heavily focused on political figures and institutions. The presence of similar political entities across all the ideological showcases that framing may differ even if the topic of discussion is the same.

5.2 Tokenization of Textual Data

In this module, the appropriate tokenization module relevant to the transformer model is loaded. This tokenization module can be Bert-base, Roberta-base or Deberta-base. The tokenizer splits the text into subword token using WordPiece tokenization. A custom tokenization function is defined on the relevant columns of the dataset; this function pads all sequences to a uniform length of 128 and truncates if the input is longer than this value. This function is then applied to the training and validation datasets.

5.3 Fine Tuning using Transformer Model

This module takes the tokenized data and a pretrained transformer model is used. The number of classes depends upon the type of classification problem is used. The dataset is split into 80:20 ratio. The model is then fine tuned on the dataset with a batch size of 4 and in 3 epochs with a learning rate of 2e-5 and weight decay is used to prevent overfitting. Weights are computed to address class imbalance. The loss function CrossEntropyLoss is computed inside the forward method only when labels are provided.

Fine-tuning Model	Dataset	Training Loss
BERT	LIAR	0.195500
BERT	UK News	0.568700
BERT	QBias	0.730800
RoBERTa	LIAR	0.259100
RoBERTa	UK News	0.651300
RoBERTa	QBias	1.100500
DeBERTa	LIAR	0.198500
DeBERTa	UK News	0.726000

DeBERTa	QBias	1.032900

Table 2. Training Loss During Fine-Tuning

The Table 2. showcases the training during fine-tuning the models on the three datasets.

5.4 MeanPooling for embedding generation

This module converts the transformer token level embeddings into a single fixed size sentence representation. Sentences can vary in length and includes padding, therefore the implements an attention mask to identify non-padded tokens. This attention mask is used to match the dimensions of the embeddings and to zero out the padded tokens and thus giving an average embedding. This operation creates a dense vector that summarizes semantic content of the input sentence. The embeddings are then generated.

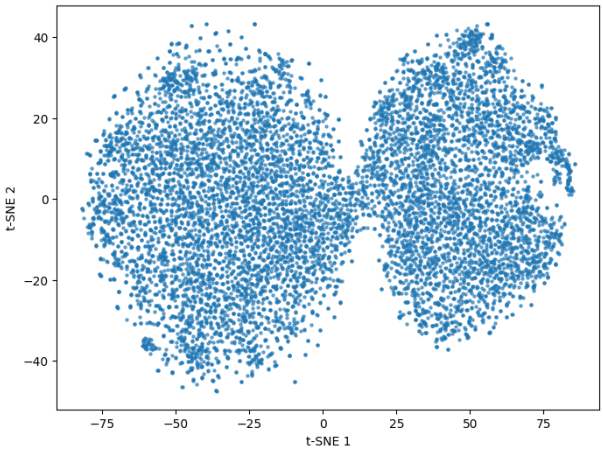


Figure 10. t-SNE BERT(LIAR)

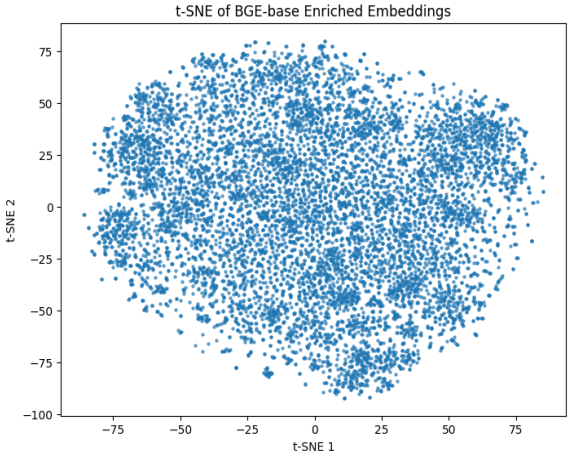


Figure 11. t-SNE BASELINE(LIAR)

Figure 10. which is t-SNE showcases that can observe that BERT embeddings were able to make much distinct clusters compared to baseline embeddings (Figure 11.).

5.5 MLP Classifier

For downstream classification, the labels depending on the type of classification and dataset is loaded using pandas and the label values are assigned to y. Meanwhile the generated embeddings are then used as input for variable X. The dataset is then split for training and testing with 80:20 ratio. Before classification the dataset which is originally in numpy array format, which is the embeddings must be converted to the appropriate Hugging Face Dataset format. Class weights are computed to give importance to minority class during training. The class weights as calculated as balanced. The model is then trained for 20 epoch with a batch size of 32.

Fine-tuning Model	Dataset	Training Loss
BERT	LIAR	0.04445
BERT	UK News	0.370078
BERT	QBias	0.52299
RoBERTa	LIAR	0.1105
RoBERTa	UK News	0.48293

RoBERTa	QBias	1.083521
DeBERTa	LIAR	0.060400
DeBERTa	UK News	0.595686
DeBERTa	QBias	1.005135

Table 3. Training Loss During Downstream Classification

The Table 3. showcases the training loss during final downstream classification on the three datasets.

5.6 Libraries and modules used

Table 4. shows the list of library and modules used for the various implementation tasks in this thesis

Library / Package	**Purpose**
transformers	Fine-tuning BERT, RoBERTa, and DeBERTa models, handling tokenization, and training setup
datasets	Creating and managing dataset objects compatible with Hugging Face pipeline
google.colab	Uploading/downloading files in Google Colab environment
sklearn.metrics	Calculating evaluation metrics such as accuracy, F1-score, confusion matrix, ROC curve, and AUC
sklearn.model_selection	Splitting data into training and validation sets
sklearn.preprocessing	Encoding class labels and binarizing labels for evaluation
torch	Core PyTorch functions for model training and tensor operations
torch.nn	Building neural network layers and defining architectures
torch.utils.data	Creating and loading datasets for training/testing
matplotlib.pyplot	Plotting graphs such as ROC curves and confusion matrices
seaborn	Creating visually appealing and informative statistical plots
numpy	Numerical computations and handling arrays (e.g., embeddings)
pandas	Data loading, preprocessing, and manipulation
tqdm	Displaying progress bars during data processing and training
os	Handling file paths and environment operations

Table 4. Libraries Used

6 Evaluation

6.1 Evaluation of Baseline Models

The Table 5. shows the performance of the baseline models on all the three datasets in terms of F1-score(macro) and Accuracy

BGE-BASE Model

Datasets	F1-Score	Accuracy
LIAR	75%	76%
UK News	43%	61%
QBias	37%	42%

E5-Base Model

Datasets	F1-Score	Accuracy
LIAR	75%	76%
UK News	41%	52%
QBias	45%	46%

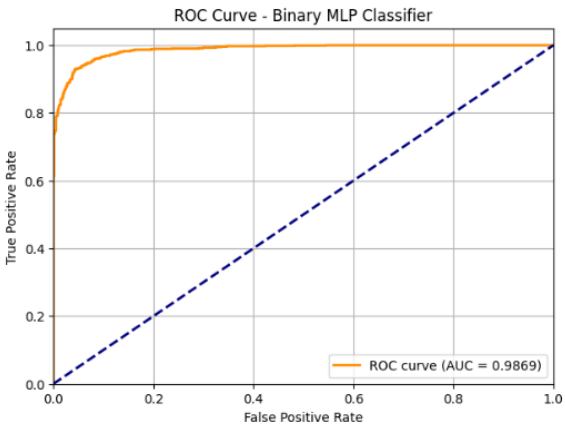
MPNET-Base Model

Datasets	F1-Score	Accuracy
LIAR	75%	76%
UK News	41%	58%
QBias	36%	39%

Table 5. Results of Base Line Models

6.2 BERT Fine Tuned embeddings Evaluation

Evaluation on LIAR DATASET:




```
Macro F1-score : 0.9380903048142317
Weighted F1-score : 0.9394455329200561
              precision    recall  f1-score   support

        left       0.91      0.95      0.93       828
       right       0.96      0.93      0.95      1133

    accuracy                           0.94      1961
   macro avg       0.94      0.94      0.94      1961
weighted avg       0.94      0.94      0.94      1961
```

Figure 12. Classification Report BERT and AUC-ROC BERT(LIAR)

From the Figure 12. it is observed that the model achieved an AUC score of 0.9869 and a macro F1-score of 0.938. Also from the classification Report it is observed that the model was able to predict both left and right wing effectively.

Evaluation on UK News Dataset

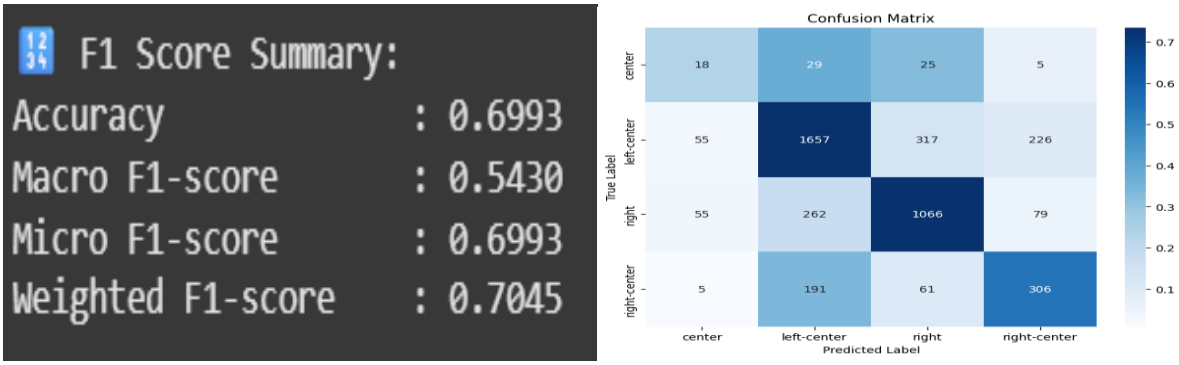


Figure 13. Performance metrics and Confusion Matrix BERT(UK NEWS)

The model achieved an accuracy of 70%, a Macro F1-score of 0.54 and a weighted F1-score of 0.70 as shown in Figure 13. From the confusion matrix it is seen that the model performs well for "left-center" and "right" but struggles with "center" and "right-center," which are often misclassified as adjacent classes.

Evaluation on QBias Dataset

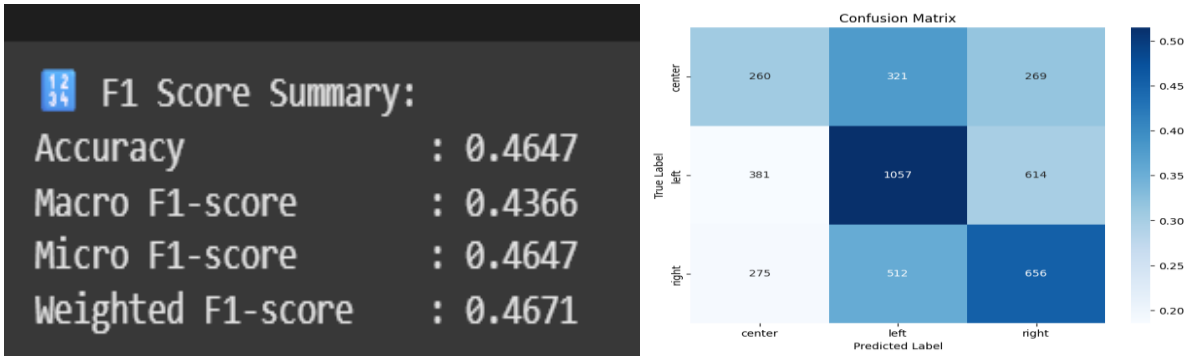


Figure 14. Performance metrics and Confusion Matrix BERT(QBias)

The model performed very poor on QBias dataset as evident in Figure 14. The Accuracy was below 50% and the Macro F1-score is 43%. And as seen in the confusion matrix, the model didn't predict any of the wing with a maximum confidence.

6.3 Roberta Fine Tuned embeddings Evaluation

Evaluation of LIAR Dataset

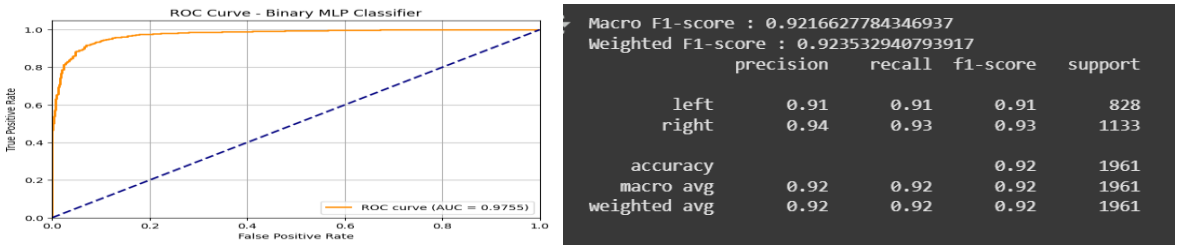


Figure 15. AUC-ROC score and Classification Report for RoBERTa(LIAR)

From the Figure 15. it is observed that the model achieved an AUC score of 0.9755 and a macro F1-score of 0.921. Also from the classification Report it is observed that the model was able to predict both left and right wing effectively with a great F1-score.

Evaluation on UK News Dataset

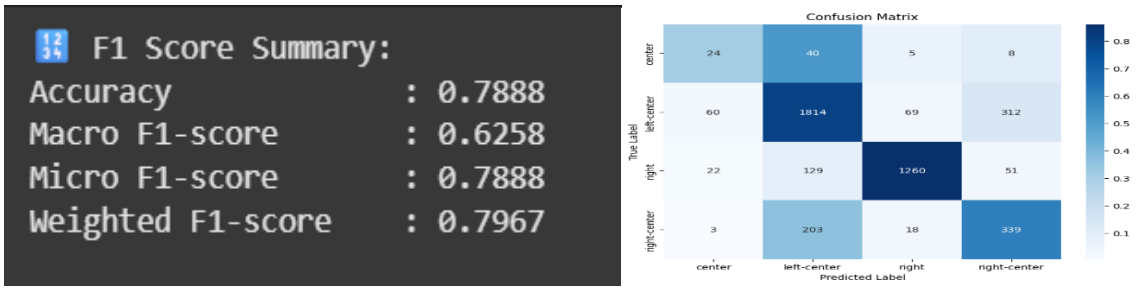


Figure 16. Performance metrics and Confusion Matrix RoBERTa(UK NEWS)

From Figure 16. The model achieved an accuracy of 78% and a Macro F1-score of 62% and a weighted F1-score of 79%. The model predicts "left-center" (1,814 correct) and "right" (1,260 correct) most accurately, while "center" (24 correct) and "right-center" (339 correct) as evident in the confusion matrix

Evaluation on QBias Dataset

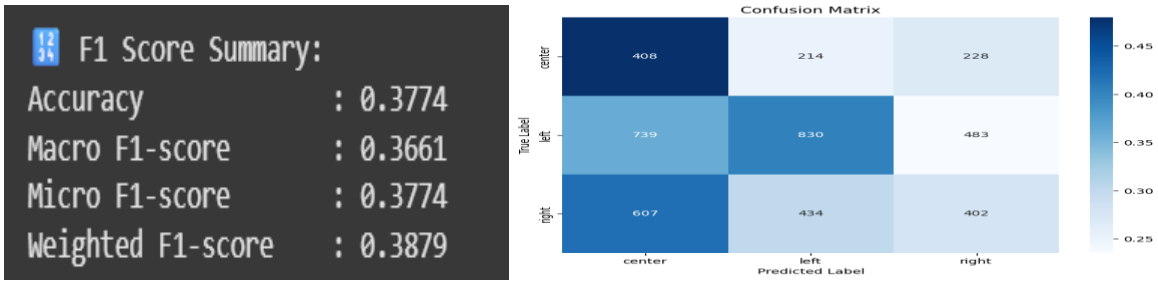


Figure 17. Performance metrics and Confusion Matrix RoBERTa(QBias)

On this dataset, as we can observe from Figure 17. The model performance was subpar with an accuracy of 37% and a macro F1-score of 36% and Weighted F1-score of 38%. The confusion matrix also indicates that the model didn't predict any of the classes accurately.

6.4 Deberta Fine Tuned Embeddings Evaluation

Evaluation on LIAR DATASET

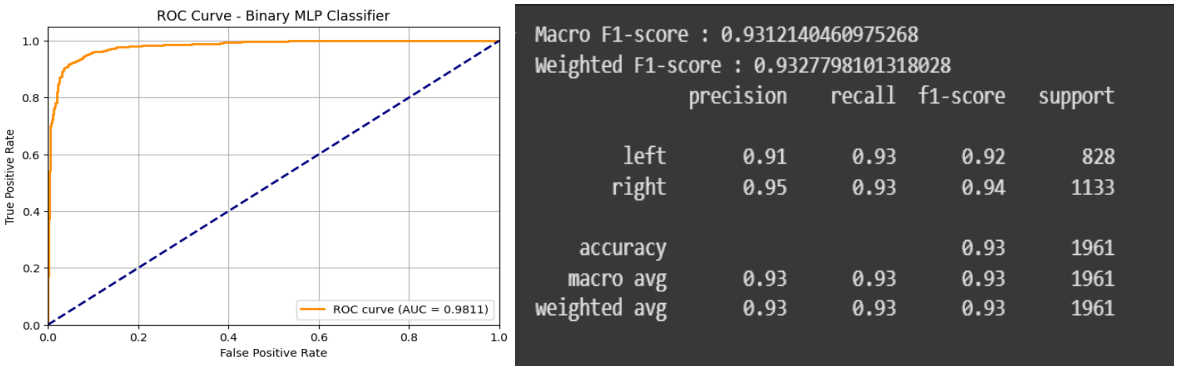


Figure 18. AUC-ROC score and Classification Report for DeBERTa(LIAR)

From the Figure 18. it is observed that the model achieved an AUC score of 0.9811 and a macro F1-score of 0.931. Also from the classification Report it is observed that the model was able to predict both left and right wing effectively.

Evaluation on UK News Dataset

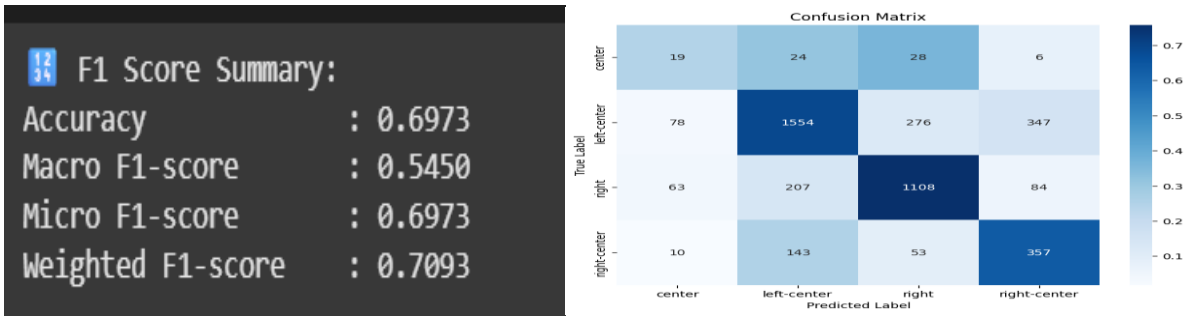


Figure 19. Performance metrics and Confusion Matrix DeBERTa(UK News)

From Figure 19. The model achieved an accuracy of 69% and a Macro F1-score of 54% and a weighted F1-score of 71%. The model performs best on "left-center" (1,554 correct) and "right" (1,108 correct) but poorly on "center" (19 correct) as shown in the confusion matrix.

Evaluation on QBias Dataset

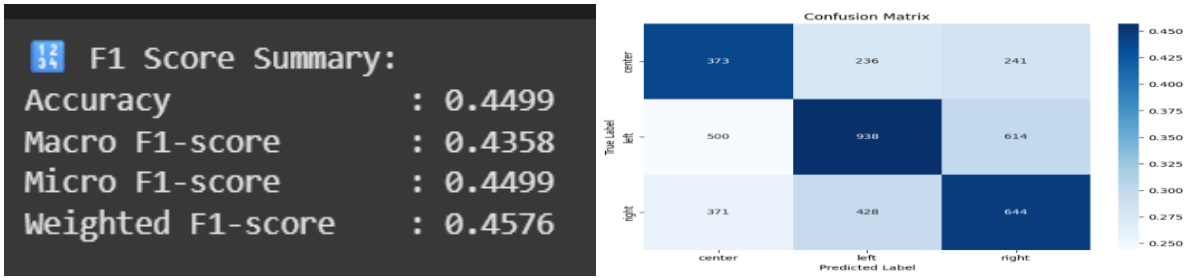


Figure 20. Performance metrics and Confusion Matrix DeBERTa(QBias)

As evident from Figure 20. It is found that the model's performance was subpar like for the other two fine-tuned models in terms of accuracy, F1-score.

The table (Table 6.) shows the individual performance of the fine-tuned models for each dataset.

Model	Dataset	Macro F1	Micro F1	Weighted F1	Accuracy	AUC (LIAR)
BERT	LIAR	94%	—	94%	94%	0.9869
	UK News	54%	70%	70%	70%	—
	QBias	44%	46%	47%	46%	—
RoBERTa	LIAR	92%	—	92%	92%	0.9755
	UK News	63%	79%	80%	79%	—
	QBias	37%	38%	39%	38%	—
DeBERTa	LIAR	93%	—	93%	93%	0.9811
	UK News	55%	70%	71%	70%	—
	QBias	44%	45%	46%	45%	—

Table 6. Evaluation of Fine-Tuned Models

6.5 Discussion

The findings across all three datasets confirm that fine-tuned transformer models far outperform baseline sentence embedding models in political bias detection. Specifically, BERT, RoBERTa, and DeBERTa, when fine-tuned over the respective dataset and then

followed by MLP-based classification, performed optimally on both binary (LIAR) as well as multi-class (UK News and QBias) tasks. Among them, BERT was outstanding with AUC of 0.9869 and macro F1-score of 94% on the LIAR dataset, showing excellent capability in distinguishing left- vs right-leaning claims.

For RQ1, we can see that Fine tuned models performed very well in terms of predicting the political bias in politically charged statements. Where BERT performed the best with a high Macro-F1 score among the other two models. This is followed by DeBERTa and RoBERTa with a miniscule degradation in performance.

Conversely for RQ2, the UK News and QBias datasets were problematic to a greater extent since they had high class imbalance. The UK News dataset had an unusually large proportion of left-center headlines, and the QBias dataset had an overwhelming proportion of left-bias articles, so it was difficult for models to generalize across minority classes such as "center" and "right." This can be seen in lower macro F1-scores, despite relatively high micro and weighted F1 performance. These scores show that although the models performed well on dominant classes, they performed poorly in identifying minority class signals, which are crucial in multi-class political bias detection. The model performance was subpar when it came to QBias dataset even though the political spectrum was smaller than that in UK News.

The embedding-based baseline models (BGE, E5, MPNET) performed reasonably well on the LIAR dataset but lagged behind on the UK News and QBias datasets. This is once again an indication of the importance of contextual awareness as well as fine-tuning, particularly when dealing with subtle or unbalanced class distributions.

The major take home here is that fine-tuned LLMs are powerful but extremely data-sensitive, class label distribution-sensitive, and sensitive to semantic overlap among classes. Even though class weighting was used in an effort to balance the class imbalances, there can be a further investigation of more complex strategies such as focal loss, dynamic sampling, or data augmentation techniques especially tailored for underclass political classes. Another is the fact that the performance difference between multi-class and binary tasks implies that ideological subtlety is still a basic challenge.

6.6 Limitations

The research, though yielding optimistic results, has been beset by a series of limitations. The most formidable was the extreme class skewness of the UK News and QBias data sets, which, despite class weighting, yielded lower macro F1-scores as well as difficulty in perceiving minority class signals. Semantic similarity among classes would also cause classification ambiguity because various political quotes shared similar topics irrespective of political ideology. Fine-tuned transformer models were in perfect working condition in ideally balanced binary issues but susceptible to biased data distributions for multi-class issues. Computational constraints of taking advantage of Google Colab's T4 GPU limited deep hyperparameter search and model scaling, which may impact maximum performance. Lastly, this thesis left out explainable AI methods for breaking down predictions, a crucial aspect in politically sensitive domains and claims.

7 Conclusion and Future Work

The project investigated the performance of fine-tuned transformer-based language models and light-weight sentence embedding models on the political bias detection task on news headlines and political statements. There were three datasets used in the research which are LIAR for binary classification (left vs right), and UK News and QBias for multi-class classification (left, center, right). The results confirms that fine-tuned models such as BERT, RoBERTa, and DeBERTa were superior over sentence embedding baselines such as MPNET, E5, and BGE, with particular focus on macro F1-score and AUC-ROC.

Of all the models tried, BERT performed very well on the LIAR dataset by achieving an AUC score of 0.9869 and a macro F1-score of 94%. UK News and QBias dataset performance was relatively poorer, between 54% to 55% macro F1-scores, due to the issue of class imbalance and subtlety of multi-class political bias. Despite class weights being used while training, the models were unable to pick up on minority class patterns, particularly in identifying "center"-aligned content. If SMOTE or Random Oversampling was used it would have increased the performance but at the cost of overfitting.

Then using these finding we can answer both our research questions;

RESEARCH QUESTION 1: How effectively can transformer based LLM models like BERT, RoBERTa and DeBERTa can predict and classify bi-partisan political biasness in politically charged claims?
The Fine Tuned Transformed based LLM models performed much better than the baseline models and even performed exceedingly well in overall terms

RESEARCH QUESTION 2: How effectively can transformer based LLM models like BERT, RoBERTa and DeBERTa can predict and classify political biasness in news headlines on a much broader political spectrum?
The Fine Tuned Transformed based LLM models performed significantly better than the baseline models but the overall performance was not great and particulary on QBias dataset the performance was subpar.

This speaks to two significant challenges for AI political bias detection: (1) the debilitating impact of class imbalance on model generalizability, and (2) the difficulty in detecting ideologically nuanced content that will not be pigeonholed into traditional left-right dichotomies. Yet the findings indicate that with the appropriate fine-tuning and feature extraction strategies, transformer models can be trustworthy instruments for real-world political bias classification tasks.

Most of the datasets were focused on USA's political environment and there is very few research and datasets done on other political environments. More research must also be done to build models that can understand the broader political spectrum much effectively. Since the nuances of statements and ideologies differ as broad the political spectrum gets. There is also a need for Explainable AI models for political bias detection. This is to analyze how the models came up with their predictions and how much relevance it is with real life analysis. This is to ensure that the models itself are not biased.

Addressing political bias is an important thing to make sure the information and content produced by the media houses are neutral. This is for overall benefits for a peaceful democratic discourse. The future work can also include to mainstream political bias detection methods and merge it into social media platforms, this is important for UN SDG-16, which ensure an inclusive society[4].

References

Abdelhameed, A.M., Mohamed, E.H. and Medhat, W. (2025) 'Detecting Bias in Arabic Political News: A Transformer-Based Ensemble Stacking Approach', in *2025 15th International Conference on Electrical Engineering (ICEENG)*. *2025 15th International Conference on Electrical Engineering (ICEENG)*, Cairo, Egypt: IEEE, pp. 1–7. Available at: https://doi.org/10.1109/ICEENG64546.2025.11031302.

Baly, R. *et al.* (2020) 'We Can Detect Your Bias: Predicting the Political Ideology of News Articles', in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online: Association for Computational Linguistics, pp. 4982–4991. Available at: https://doi.org/10.18653/v1/2020.emnlp-main.404.

Gangula, R.R.R., Duggenpudi, S.R. and Mamidi, R. (2019) 'Detecting Political Bias in News Articles Using Headline Attention', in *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, Florence, Italy: Association for Computational Linguistics, pp. 77–84. Available at: https://doi.org/10.18653/v1/W19-4809.

Gunhal, P. *et al.* (2022) 'Stance Detection of Political Tweets with Transformer Architectures', in *2022 13th International Conference on Information and Communication Technology Convergence (ICTC)*. *2022 13th International Conference on Information and Communication Technology Convergence (ICTC)*, Jeju Island, Korea, Republic of: IEEE, pp. 658–663. Available at: https://doi.org/10.1109/ICTC55196.2022.9952951.

Haak, F. and Schaer, P. (2023) 'Qbias - A Dataset on Media Bias in Search Queries and Query Suggestions', in *Proceedings of the 15th ACM Web Science Conference 2023*. *WebSci '23: 15th ACM Web Science Conference 2023*, Austin TX USA: ACM, pp. 239–244. Available at: https://doi.org/10.1145/3578503.3583628.

Iyyer, M. *et al.* (2014) 'Political Ideology Detection Using Recursive Neural Networks', in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Baltimore, Maryland: Association for Computational Linguistics, pp. 1113–1122. Available at: https://doi.org/10.3115/v1/P14-1105.

[4] United Nations (n.d.). *The 17 sustainable development goals*. [online] United Nations. Available at: https://sdgs.un.org/goals.

Kim, K.-M. *et al.* (2023) ‘Multi-Stage Prompt Tuning for Political Perspective Detection in Low-Resource Settings’, *Applied Sciences*, 13(10), p. 6252. Available at: https://doi.org/10.3390/app13106252.

Krieger, J.-D. *et al.* (2022) ‘A domain-adaptive pre-training approach for language bias detection in news’, in *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*. *JCDL ’22: The ACM/IEEE Joint Conference on Digital Libraries in 2022*, Cologne Germany: ACM, pp. 1–7. Available at: https://doi.org/10.1145/3529372.3530932.

Lunardi, R., La Barbera, D. and Roitero, K. (2024) ‘The Elusiveness of Detecting Political Bias in Language Models’, in *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*. *CIKM ’24: The 33rd ACM International Conference on Information and Knowledge Management*, Boise ID USA: ACM, pp. 3922–3926. Available at: https://doi.org/10.1145/3627673.3680002.

Luo, K. *et al.* (2024) ‘BGE Landmark Embedding: A Chunking-Free Embedding Method For Retrieval Augmented Long-Context Large Language Models’. arXiv. Available at: https://doi.org/10.48550/ARXIV.2402.11573.

Menzner, T. and Leidner, J.L. (2024) ‘Improved Models for Media Bias Detection and Subcategorization’, in A. Rapp et al. (eds) *Natural Language Processing and Information Systems*. Cham: Springer Nature Switzerland (Lecture Notes in Computer Science), pp. 181–196. Available at: https://doi.org/10.1007/978-3-031-70239-6_13.

Pavlov, T. and Mirceva, G. (2022) ‘COVID-19 Fake News Detection by Using BERT and RoBERTa models’, in *2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO)*. *2022 45th Jubilee International Convention on Information, Communication and Electronic Technology (MIPRO)*, Opatija, Croatia: IEEE, pp. 312–316. Available at: https://doi.org/10.23919/MIPRO55190.2022.9803414.

Qazi, M., Khan, M.U.S. and Ali, M. (2020) ‘Detection of Fake News Using Transformer Model’, in *2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*. *2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, Sukkur, Pakistan: IEEE, pp. 1–6. Available at: https://doi.org/10.1109/iCoMET48670.2020.9074071.

Rönnback, R., Emmery, C. and Brighton, H. (2025) ‘Automatic large-scale political bias detection of news outlets’, *PLOS One*. Edited by S. Elbassuoni, 20(5), p. e0321418. Available at: https://doi.org/10.1371/journal.pone.0321418.

Song, K. *et al.* (2020) ‘MPNet: Masked and Permuted Pre-training for Language Understanding’. arXiv. Available at: https://doi.org/10.48550/arXiv.2004.09297.

V, R., Singh, A. and Kumar, K. (2025) ‘Measuring Political Bias in LLMs using Fine-Tuning RoBERTa Model’, in *2025 3rd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)*. *2025 3rd International Conference on Self Sustainable Artificial Intelligence Systems (ICSSAS)*, Erode, India: IEEE, pp. 1213–1218. Available at: https://doi.org/10.1109/ICSSAS66150.2025.11081114.

Wang, L. *et al.* (2024) 'Text Embeddings by Weakly-Supervised Contrastive Pre-training'. arXiv. Available at: https://doi.org/10.48550/arXiv.2212.03533.

Wang, W.Y. (2017) '"Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection', in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Vancouver, Canada: Association for Computational Linguistics, pp. 422–426. Available at: https://doi.org/10.18653/v1/P17-2067.

Won-Tae Joo, Jeong, Y.-S., and KyoJoong Oh (2016) 'Political orientation detection on Korean newspapers via sentence embedding and deep learning', in *2016 International Conference on Big Data and Smart Computing (BigComp)*. *2016 International Conference on Big Data and Smart Computing (BigComp)*, Hong Kong, China: IEEE, pp. 502–504. Available at: https://doi.org/10.1109/BIGCOMP.2016.7425979.

Zhu, J., Cukier, M. and Cruickshank, I. (2025) 'DocNet: Semantic Structure in Inductive Bias Detection Models', in *Proceedings of the 17th ACM Web Science Conference 2025*. *Websci '25: 17th ACM Web Science Conference 2025*, New Brunswick NJ USA: ACM, pp. 138–147. Available at: https://doi.org/10.1145/3717867.3717889.